Semantische Nachhaltigkeit und Kontrolle: Gedanken zu schema.org und Linked Open Data[1]

1 Linked Open Data: ein fröhliches Chaos

> "It's massively successful. It is trivially simple. Massively successful like karaoke - anybody can do it." ((Nelson, 2001))
>
> "Anyone can say anything about anything" (Berners Lee, 2002, according to (Klyne & Caroll, 2002))

Die Entwicklung des WWW von einem ersten Entwurf als Web der Dokumente über die zu Anfang jedenfalls von den Allmachtsphantasien der „Künstlichen Intelligenz" geprägte Phase des 'Semantic Web' bis zu seiner heutigen Ausprägung als Web der Dinge, als 'Linked Data Web' ist auf den ersten Blick eine unglaubliche Erfolgsgeschichte: hier ist in wenig mehr als 20 Jahren eine gigantische, weltumspannende und komplexvernetzte Maschine für die Wissensgenerierung entstanden. Vielleicht am besten charakterisiert ist das entstandene Resultat durch den Titel von (Heath & Bizer, 2011): das WWW ist inzwischen in der Tat zu einem „Global Data Space" geworden.

Dabei waren in der Entwicklung vom 'Document Web' der ersten Generation zum heutigen WWW zwei elementare Erweiterungen maßgeblich: die syntaktische Erweiterung der WWW-Protokolle um das Resource Description Framework (RDF) erlaubte die semantische Typisierung von Verbindungen zwischen WWW-Entitäten und damit – zusammen mit der Schemasprache RDFS – eine maschinelle Verarbeitung von WWW-Inhalten im Sinne einfacher, deterministischer Schlußoperationen, und eine Erweiterung im Repräsentationsraum des WWW erlaubt nunmehr die Repräsentation im Prinzip beliebiger Entitäten der 'wirklichen' und der gedachten Welt im WWW.

Dabei sind drei Charakteristika bemerkenswert, die im Folgenden polemisch betrachtet werden sollen:

- Es gibt keine 'Regierung' des WWW, welche die Macht hätte, über opportune bzw. nicht genehme Äußerungen zu bestimmen. Das Motto vorangestellte Zitat von Berners-Lee sagt es sehr treffend: jede Äußerung zu jedem Thema ist im WWW für jedermann möglich, eine Zensur findet nicht statt und ist auch so gut wie nicht implementierbar. Staaten wie China, denen dies ein Dorn im Auge ist, haben keine andere Wahl, als ihre gesamte Population hinter einer riesigen Proxyarchitektur zu verstecken und sie damit aus dem WWW zu verbannen: im WWW wäre sie nicht effektiv kontrollierbar.
- Das massive Wachstum des Linked Data Web seit 2007 wurde technisch ermöglicht vor allem durch eine massive Vereinfachung des komplexen Schichtenmodells des 'Semantic Web' und eine Rückbesinnung auf RDF als zentralen technischen Ankerpunkt sowie darauf basierende robuste technische Standards des W3C wie etwa SPARQL.

1 Ich habe an anderer Stelle (so zuletzt in (Gradmann, 2013)) die Entwicklung des 'Semantic Web' und des 'Linked Data' Paradigmas überblicksartig nachgezeichnet, so dass an dieser Stelle ein erneutes Referieren dieser Entwicklung verzichtbar ist: die nachstehenden Ausführungen mögen vielmehr als Vertiefung von Aspekten gelesen werden, die dort ebenfalls schon angerissen sind. Weiter verweise ich in diesem Zusammenhang auf die drei grundlegenden, lehrbuchartigen Einführungen: (Antoniou & Van Harmelen, 2008), (Hitzler, Krötzsch, Rudolph, & Sure, 2008) und (Heath & Bizer, 2011). Alle drei sind auch für eine vertiefende Einführung empfehlenswert.

- Technische Vereinfachung und die urdemokratische Entwicklung des Web als Aussagensystem mitsamt der damit verbundenen Freiheit, Ontologieressourcen frei und unkontrolliert zu kreieren und zu verwenden haben ihren Preis: der Blick auf eine URL wie http://ws.nju.edu.cn/falcons/ontologysearch/result.jsp?query=person offenbart ein massives Qualitätsproblem des Linked Data Web, denn die Ontologiesuchmaschine Falcons macht dort sichtbar, dass es an Ontologieressourcen für die Modellierung etwa von Personen ganz sicher nicht mangelt: es sind deren im Gegenteil deutlich zu viele.

Das W3C ist sich gerade des letzten Problems durchaus bewusst und hat es lange Zeit auch bewusst toleriert, wie das folgende Zitat von (Klyne & Caroll, 2002) belegt:

> „To facilitate operation at Internet scale, RDF is an open-world framework that allows anyone to say anything about anything. In general, it is not assumed that all information about any topic is available. A consequence of this is that RDF cannot prevent anyone from making nonsensical or inconsistent assertions, and applications that build upon RDF must find ways to deal with conflicting sources of information. (This is where RDF departs from the XML approach to data representation, which is generally quite prescriptive and aims to present an application with information that is well-formed and complete for the application's needs.)"

Neben den anderen 'Großbaustellen' des WWW (fehlende Versionierungsmethoden, mangelnde Expressivität in den Bereichen Provenienz und Autorisierungsmerkmale) hat das Linked Data Web also ein massives Qualitätsproblem vor allem in Gestalt der vielen koexistierenden und unkontrollierten Ontologien, die einerseits zum Teil enorme Überschneidungsbereiche haben, andererseits aber doch immer nur partiell semantisch redundant sind, so daß ein Konstrukt wie owl:sameAs streng genommen so gut wie nie korrekt anwendbar ist.

Damit ähnelt das fröhliche Chaos des Linked Data Web – wollte man eine architektonische Metapher gebrauchen – am ehesten einer dynamisch wuchernden brasilianischen Favela, und es beginnt in der 'Community' das Nachdenken darüber, wie dies Qualitätsproblem anzugehen wäre: Ontology-Mapping sowie der Bewertung und Modellierung semantischer Nähe und Überschneidung sind zwei der momentan aktivsten Forschungsfelder in diesem Umfeld.

2 Schema.org: Ordnung, Disziplin und Kontrolle

Ganz anders als die Favela-artige Herangehensweise mutet die Strategie der schema.org-Initiative an, die – wieder in einer architektonischen Metapher – weit eher dem wohlorganisierten, geordneten Bau einer Kathedrale ähnelt.

Lanciert wurde diese Initiative am 02. Juni 2011 von den Suchmaschinen Google, Bing (Microsoft) und Yahoo, einige Monate später kam die führende russische Suchmaschine Yandex dazu. Das Ziel von schema.org wird im Blog der Initiative wie folgt benannt: "create and support a standard set of schemas for structured data markup on web pages" (O'Connor, 2011)

Es geht also um den Aufbau eines kontrolliertes Kernvokabular, der zwar Erweiterungen durchaus zulässt, jedoch immerhin eine zentrale Redaktionsinstanz kennt. Neben diesem Charakteristikum unterschied schema.org sich zumindest im ersten Angang noch in einem zweiten Punkt entscheidend vom Linked Data Web: die Aussagensyntax von schema.org war ursprünglich auf microdata beschränkt, RDFa kam im September 2011 dazu: im ersten Ansatz war schema.org also bewusst nicht als Teil des 'offenen' WWW entworfen.

Dabei unterliegt der Weltsicht von schema.org, wie sie unter http://schema.org/docs/full.html dokumentiert ist, einer eigenartigen Verzerrung, vergleicht man sie mit derjenigen von Top-Level-Ontologien wie etwa SUMO[2] und der darunter liegenden Mid-Level Ontology MILO. Deutlich überrepräsentiert sind hier – wen ig verwunderlich! - die Personengruppen und Institutionen/Firmen, die das klassische, auf Werbeeinnahmen basierende Geschäftsmodell der großen Suchmaschinen befeuern. Und kaum prominent sind Entitäten, mit denen sich im WWW kein Geld verdienen lässt. So nehmen etwa Diätpläne („diet") in der Klasse „Creative Work" als direkte Subklasse eine deutlich prominentere Rolle ein, als die zwei Hierarchiestufen tiefer angesiedelten wissenschaftlichen Aufsätze („scholarly article"), die dann außerdem kurioserweise noch als eigene Unterklasse „medical articles" gesondert behandeln – auch hier ist wahrscheinlich wieder eine monetär geprägte Perspektive im Spiel. Oder – um ein anderes Beispiel zu nennen – es sind innerhalb von „Creative Work" gleich drei eigene Unterklassen für „TVEpisode", „TVSeason" und „TVSeries" vorgesehen.

Dennoch muß zugunsten von schema.org festgehalten werden, daß die Ontologie das selbstgesetzte Ziel eines einheitlichen Standardvokabulars ohne die massiven Redundanzen des Linked Data Web zumindest ein Stück weit erreicht: das Schema http://schema.org/Person jedenfalls wirkt deutlich aufgeräumter als das oben angeführte Beispiel der Personenontologien im Linked Data Web – wenngleich auch hier wieder die kommerzielle Verzerrung des Suchmschinen-Geschäftsmodells durchschlägt, denn unter den ansonsten eher generischen Attributen von „Person" finden sich völlig isoliert und überraschen die zwei sehr spezifischen Attribute „hasPOS" und „brand" ...

Ähnlich steht es mit schema.org basierten Anwendungen: das hierfür generell als Quelle interessante Verzeichnis http://linter.structured-data.org/examples/ führt ganz überwiegend Verzeichnis- und Vertriebsdienste aus dem Bereich der Kreativindustrie auf.

Vor allem aber fällt auf, wie wenig Verbindung die Vokabularschemata von schema.org mit den Ontologien des Linked Data Web haben, modellieren sie doch durchweg Entitäten, die selbst schon eine Repräsentation im Linked Data Web haben. Schema.org fügt also den Vokabularressourcen im WWW eine weitere, nahezu 100% semantisch redundante Vokabularschicht hinzu, was unmittelbar zur Frage nach der Motivation der Initiatoren von schema.org führt: warum diese großangelegte, hochgradig redundante Initiative?

3 Eine versteckte Agenda?

Von Beginn an war die Diskussion über schema.org im WWW von heftigen Spekulationen und mitunter geradezu paranoiden Untertönen geprägt, genährt allerfdings nicht zuletzt durch die notorische Intransparenz von Google: die Firma ist bis heute Antworten auf eine ganze Reihe sich unmittelbar aufdrängender Fragen schuldig geblieben.

So ist etwa bis heute unklar, was die harten Konkurrenten Google, Microsoft und Yahoo dazu bewogen hat, ausgerechnet im Bereich ihres Suchmaschinen-Kerngeschäfts zu kooperieren. Handelt es sich womöglich um eine Reaktion auf das als bedrohlich wahrgenommene Funktionsmodell des Linked-Data-Web in der Absicht, Kunden möglichst nachhaltig an das Suchmaschinen-Geschäftsmodell zu binden?

Und auch die Motivation der ursprünglichen Einengung auf die Microdata-Syntax und die damit verbundene Ablehnung von RDF ist niemals wirklich deutlich geworden, zu-

2 S. http://www.ontologyportal.org/

mal diese Linie dann schon nach wenigen Monaten verlassen wurde: warum hat man sich nicht von vorne herein auf das RDF-Paradigma eingelassen?

Und schließlich: gibt es einen Zusammenhang zwischen schema.org und dem wenig später lancierten Knowledge Graph, der wie eine Google-spezifische Antwort auf das Linked-Data-Web daherkommt, und wenn ja, welcher Art ist dieser Zusammenhang? Manu Sporny hat zu diesem Thema in (Sporny, 2012) interessante Details zusammengetragen.

Genereller gesprochen: ist das Verhältnis der Paradigmen schema.org und Linked Open Data in Analogie zu begreifen zu der metaphorischen Opposition The Cathedral vs. The Bazaar, wie sie in (Raymond, 2001) für die Opposition Open Source vs. Closed Source eingeführt worden war? Ist schema.org also nützlich, oder zumindest unschuldig, oder womöglich gar böse?

Letzterer Position scheint (Stewart, 2011) zuzuneigen, wenn er in seinem Blog schreibt:

> "Schema.org appears to be Linked Data Lite with extremely limited support for vocabularies outside of the service. […] There is a subtle air of intimidation throughout the schema.org announcements and documentation. […] Again, I could just be paranoid, but this is Microsoft and Google we're talking about. Whatever happened to "do no evil?""

Und selbst wenn man so weit nicht gehen möchte und jeden paranoiden Denkreflex vermeidet bleibt die Eingangsfrage unbeantwortet, was Google, Microsoft und Yahoo dazu bewogen hat, eine eigene, semantisch redundante Initiative zu lancieren, die das Linked Data Web im Grunde noch einmal zu erfinden versucht.

4 Semantischer Darwinismus - oder doch Zensur??

Die Antwort könnte – vordergründig betrachtet – in zwei alternativen Annahmen bestehen. Entweder handelt es sich bei schema.org um den zutiefst unschuldigen, raubtierkapitalistischen Versuch, eine große, aber diffuse semantische Ressource (das Linked Data Web) durch eine andere, bessere Lösung zu ersetzen und damit nur um eine radikale Variante dessen was wir möglicherweise bald 'ontologischen Darwinismus' nennen könnten. Oder aber es handelt sich um eine Art von Outsourcingmodell für Zensur, dass letztendlich darauf abzielt, nur noch diejenigen Inhalte im WWW effektiv präsent werden zu lassen, die das 'richtige' Vokabular verwenden – nur dass dann eben die Zensurhoheit nicht mehr bei politisch-administrativen Instanzen liegt, wie in der Vergangenheit, sondern bei Google & co.

Vielleicht jedoch besteht die richtige Antwort in einer sehr viel weiter gehenden Annahme, die in gewisser Hinsicht substantiell über den traditionellen Zensurbegriff hinausgeht. Denn die traditionelle Zensur war eine Methode, Menschen an der Veröffentlichung ihres Schaffens zu hindern, nicht jedoch an der geistigen Produktion selbst. Insbesondere waren die Sprache und sonstigen künstlerischen Hilfsmittel dabei niemandes partikulares Eigentum – dies könnte sich, zumindest im Sinne der Wahrnehmbarkeit von Aussagen im WWW, mit schema.org in dem Sinne ändern, dass damit die Ausdrucksmittel selbst oligopolistisches Eigentum des Suchmaschinenkonsortiums werden. Wieweit die Zusammenarbeit des schema.org-Konsortiums mit dem W3C diese Annahme relativieren kann ist aus meiner Sicht momentan nicht einschätzbar.

In einer sehr pessimistischen Lesart ginge es dann also bei schema.org nicht nur darum, was gesagt werden kann (bzw. nicht gesagt werden kann), und wer das zugrunde liegende Vokabular kontrolliert, sondern vor allem darum, welche Aussagen überhaupt

effektiv wahrgenommen werden (und welche einfach nicht registriert werden). Und damit ginge es dann bei schema.org gar nicht primär um Geld, sondern um die Leitwährung des WWW: um Aufmerksamkeit!

5 Literatur

Antoniou, G., & Van Harmelen, F. (2008). *A Semantic Web Primer* (2nd Edition.). MIT-Press. Retrieved from http://dl.acm.org/citation.cfm?id=1373341

Gradmann, S. (2013). Semantic Web und Linked Open Data. In *Grundlagen der praktischen Information und Dokumentation.* (6. Ausgabe., pp. 219 – 228). Berlin: de Gruyter.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool. Retrieved from http://linkeddatabook.com/editions/1.0/

Hitzler, P., Krötzsch, M., Rudolph, S., & Sure, Y. (2008). *Semantic-Web-Grundlagen*. Heidelberg: Springer. Retrieved from http://www.semantic-web-grundlagen.de/

Klyne, G., & Caroll, J. (2002, August 29). Resource Description Framework (RDF): Concepts and Abstract Data Model. W3C. Retrieved from http://www.w3.org/TR/2002/WD-rdf-concepts-20020829/

Nelson, T. (2001, October 8). Visionary lays into the web. Retrieved from http://news.bbc.co.uk/2/hi/science/nature/1581891.stm

O'Connor, M. (2011, July 20). schema blog. Retrieved from http://blog.schema.org/search?updated-max=2011-12-12T12:10:00-08:00

Raymond, E. S. (2001). *The Cathedral & the Bazaar*. Bejing etc.: O'Reilly Media.

Sporny, M. (2012, February 14). Google Indexing RDFa 1.0 + schema.org Markup. *The Beautiful, Tormented Machine*. Retrieved from http://manu.sporny.org/2012/google-indexing-schema-rdfa/

Stewart, D. (2011, June 4). Schema.org: Webmaster One-Stop or Linked Data Land Grab? *Gartner Blog*. Retrieved from http://blogs.gartner.com/darin-stewart/2011/06/04/schema-org-webmaster-one-stop-or-linked-data-land-grab/